1

Introduction

For humans, speech is the quickest and most natural form of communication. Beginning in the late 19th century, verbal communication has been systematically extended through technologies such as radio broadcast, telephony, TV, CD and MP3 players, mobile phones and the Internet by voice over IP. In addition to these examples of one and two way verbal human–human interaction, in the last decades, a great deal of research has been devoted to extending our capacity of verbal communication with computers through *automatic speech recognition* (ASR) and speech synthesis. The goal of this research effort has been and remains to enable simple and natural *human–computer interaction* (HCI). Achieving this goal is of paramount importance, as verbal communication is not only fast and convenient, but also the only feasible means of HCI in a broad variety of circumstances. For example, while driving, it is much safer to simply ask a car navigation system for directions, and to receive them verbally, than to use a keyboard for tactile input and a screen for visual feedback. Moreover, hands-free computing is also accessible for disabled users.

1.1 Research and Applications in Academia and Industry

Hands-free computing, much like hands-free speech processing, refers to computer interface configurations which allow an interaction between the human user and computer without the use of the hands. Specifically, this implies that no close-talking microphone is required. Hands-free computing is important because it is useful in a broad variety of applications where the use of other common interface devices, such as a mouse or keyboard, are impractical or impossible. Examples of some currently available hands-free computing devices are camera-based head location and orientation-tracking systems, as well as gesture-tracking systems. Of the various hands-free input modalities, however, *distant speech recognition* (DSR) systems provide by far the most flexibility. When used in combination with other hands-free modalities, they provide for a broad variety of HCI possibilities. For example, in combination with a pointing gesture system it would become possible to turn on a particular light in the room by pointing at it while saying, "Turn on this light."

The remainder of this section describes a variety of applications where speech recognition technology is currently under development or already available commercially. The

Distant Speech Recognition Matthias Wölfel and John McDonough

application areas include intelligent home and office environments, humanoid robots, automobiles, and speech-to-speech translation.

1.1.1 Intelligent Home and Office Environments

A great deal of research effort is directed towards equipping household and office devices – such as appliances, entertainment centers, personal digital assistants and computers, phones or lights – with more user friendly interfaces. These devices should be unobtrusive and should not require any special attention from the user. Ideally such devices should know the mental state of the user and act accordingly, gradually relieving household inhabitants and office workers from the chore of manual control of the environment. This is possible only through the application of sophisticated algorithms such as speech and speaker recognition applied to data captured with far-field sensors.

In addition to applications centered on HCI, computers are gradually gaining the capacity of acting as mediators for human–human interaction. The goal of the research in this area is to build a computer that will serve human users in their interactions with other human users; instead of requiring that users concentrate on their interactions with the machine itself, the machine will provide ancillary services enabling users to attend exclusively to their interactions with other people. Based on a detailed understanding of human perceptual context, intelligent rooms will be able to provide active assistance without any explicit request from the users, thereby requiring a minimum of attention from and creating no interruptions for their human users. In addition to speech recognition, such services need qualitative human analysis and human factors, natural scene analysis, multimodal structure and content analysis, and HCI. All of these capabilities must also be integrated into a single system.

Such interaction scenarios have been addressed by the recent projects *Computers in the Human Interaction Loop* (CHIL), *Augmented Multi-party Interaction* (AMI), as well as the successor of the latter *Augmented Multi-party Interaction with Distance Access* (AMIDA), all of which were sponsored by the European Union. To provide such services requires technology that models human users, their activities, and intentions. Automatically recognizing and understanding human speech plays a fundamental role in developing such technology. Therefore, all of the projects mentioned above have sought to develop technology for automatic transcription using speech data captured with distant microphones, determining who spoke when and where, and providing other useful services such as the summarizations of verbal dialogues. Similarly, the *Cognitive Assistant that Learns and Organizes* (CALO) project sponsored by the US *Defense Advanced Research Project Agency* (DARPA), takes as its goal the extraction of information from audio data captured during group interactions.

A typical meeting scenario as addressed by the AMIDA project is shown in Figure 1.1. Note the three microphone arrays placed at various locations on the table, which are intended to capture far-field speech for speaker tracking, beamforming, and DSR experiments. Although not shown in the photograph, the meeting participants typically also wear close-talking microphones to provide the best possible sound capture as a reference against which to judge the performance of the DSR system.

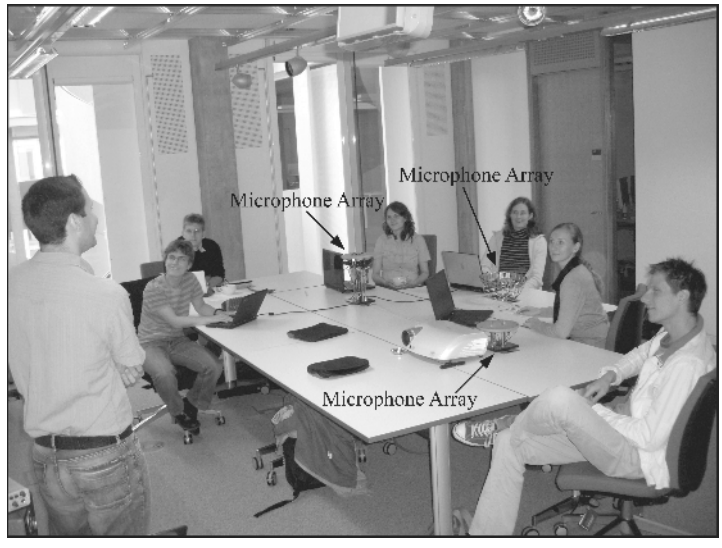

Figure 1.1 A typical AMIDA interaction. (© Photo reproduced by permission of the University of Edinburgh)

1.1.2 Humanoid Robots

If humanoid robots are ever to be accepted as full 'partners' by their human users, they must eventually develop perceptual capabilities similar to those possessed by humans, as well as the capacity of performing a diverse collection of tasks, including learning, reasoning, communicating and forming goals through interaction with both users and instructors. To provide for such capabilities, ASR is essential, because, as mentioned previously, spoken communication is the most common and flexible form of communication between people. To provide a natural interaction between a human and a humanoid robot requires not only the development of speech recognition systems capable of functioning reliably on data captured with far-field sensors, but also natural language capabilities including a sense of social interrelations and hierarchies.

In recent years, humanoid robots, albeit with very limited capabilities, have become commonplace. They are, for example, deployed as entertainment or information systems. Figure 1.2 shows an example of such a robot, namely, the humanoid tour guide robot *TPR-Robina*[1] developed by Toyota. The robot is able to escort visitors around the Toyota Kaikan Exhibition Hall and to interact with them through a combination of verbal communication and gestures.

While humanoid robots programmed for a limited range of tasks are already in widespread use, such systems lack the capability of learning and adapting to new environments. The development of such a capacity is essential for humanoid robots to become helpful in everyday life. The *Cognitive Systems for Cognitive Assistants* (COSY) project, financed by the European Union, has the objective to develops two kind of robots providing such advanced capabilities. The first robot will find its way around a

[1] ROBINA stands for ROBot as INtelligent Assistant.

Figure 1.2 Humanoid tour guide robot TPR-Robina by Toyota which escort visitors around Toyota Kaikan Exhibition Hall in Toyota City, Aichi Prefecture, Japan. (© Photo reproduced by permission of Toyota Motor Corporation)

complex building, showing others where to go and answering questions about routes and locations. The second will be able to manipulate structured objects on a table top. A photograph of the second COSY robot during an interaction session is shown in Figure 1.3.

1.1.3 Automobiles

There is a growing trend in the automotive industry towards increasing both the number and the complexity of the features available in high end models. Such features include entertainment, navigation, and telematics systems, all of which compete for the driver's visual and auditory attention, and can increase his cognitive load. ASR in such automobile environments would promote the "Eyes on the road, hands on the wheel" philosophy. This would not only provide more convenience for the driver, but would in addition actually

Figure 1.3 Humanoid robot under development for the COSY project. (© Photo reproduced by permission of DFKI GmbH)

enhance automotive safety. The enhanced safety is provided by hands-free operation of everything but the car itself and thus would leave the driver free to concentrate on the road and the traffic. Most luxury cars already have some sort of voice-control system which are, for example, able to provide

- *voice-activated, hands-free calling*
  Allows anyone in the contact list of the driver's mobile phone to be called by voice command.
- *voice-activated music*
  Enables browsing through music using voice commands.
- *audible information and text messages*
  Makes it possible to synthesize information and text messages, and have them read out loud through speech synthesis.

This and other voice-controlled functionality will become available in the mass market in the near future. An example of a voice-controlled car navigation system is shown in Figure 1.4.

While high-end consumer automobiles have ever more features available, all of which represent potential distractions from the task of driving the car, a police automobile has far more devices that place demands on the driver's attention. The goal of Project54 is to measure the cognitive load of New Hampshire state policeman – who are using speech-based interfaces in their cars – during the course of their duties. Shown in Figure 1.5 is the car simulator used by Project54 to measure the response times of police officers when confronted with the task of driving a police cruiser as well as manipulating the several devices contained therein through a speech interface.

Figure 1.4 Voice-controlled car navigation system by Becker. (© Photo reproduced by permission of Herman/Becker Automotive Systems GmbH)

Figure 1.5 Automobile simulator at the University of New Hampshire. (© Photo reproduced by permission of University of New Hampshire)

1.1.4 Speech-to-Speech Translation

Speech-to-speech translation systems provide a platform enabling communication with others without the requirement of speaking or understanding a common language. Given the nearly 6,000 different languages presently spoken somewhere on the Earth, and the ever-increasing rate of globalization and frequency of travel, this is a capacity that will in future be ever more in demand.

Even though speech-to-speech translation remains a very challenging task, commercial products are already available that enable meaningful interactions in several scenarios. One such system from *National Telephone and Telegraph* (NTT) DoCoMo of Japan works on a common cell phone, as shown in Figure 1.6, providing voice-activated Japanese–English and Japanese–Chinese translation. In a typical interaction, the user speaks short Japanese phrases or sentences into the mobile phone. As the mobile phone does not provide enough computational power for complete speech-to-text translation, the speech signal is transformed into enhanced speech features which are transmitted to a server. The server, operated by ATR-Trek, recognizes the speech and provides statistical translations, which are then displayed on the screen of the cell-phone. The current system works for both Japanese–English and Japanese–Chinese language pairs, offering translation in

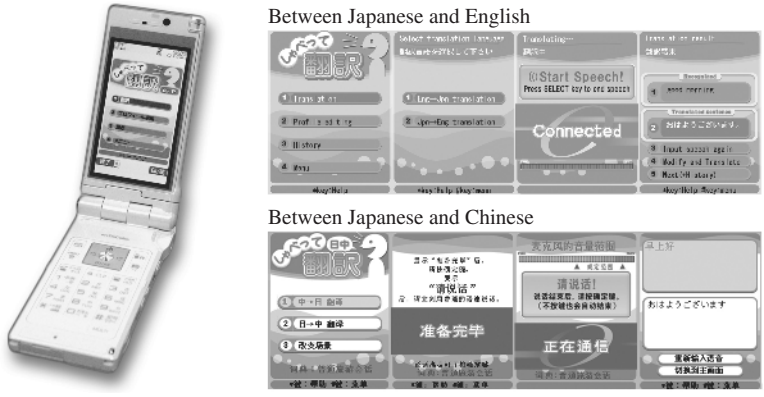

Figure 1.6 Cell phone, 905i Series by NTT DoCoMo, providing speech translation between English and Japanese, and Chinese and Japanese developed by ATR and ATR-Trek. This service is commercially available from NTT DoCoMo. (© Photos reproduced by permission of ATR-Trek)

both directions. For the future, however, preparation is underway to include support for additional languages.

As the translations appear on the screen of the cell phone in the DoCoMo system, there is a natural desire by users to hold the phone so that the screen is visible instead of next to the ear. This would imply that the microphone is no longer only a few centimeters from the mouth; i.e., we would have once more a distant speech recognition scenario. Indeed, there is a similar trend in all hand-held devices supporting speech input.

Accurate translation of unrestricted speech is well beyond the capability of today's state-of-the-art research systems. Therefore, advances are needed to improve the technologies for both speech recognition and speech translation. The development of such technologies are the goals of the *Technology and Corpora for Speech-to-Speech Translation* (TC-Star) project, financially supported by European Union, as well as the *Global Autonomous Language Exploitation* (GALE) project sponsored by the DARPA. These projects respectively aim to develop the capability for unconstrained conversational speech-to-speech translation of English speeches given in the European Parliament, and of broadcast news in Chinese or Arabic.

1.2 Challenges in Distant Speech Recognition

To guarantee high-quality sound capture, the microphones used in an ASR system should be located at a fixed position, very close to the sound source, namely, the mouth of the speaker. Thus body mounted microphones, such as head-sets or lapel microphones, provide the highest sound quality. Such microphones are not practical in a broad variety of situations, however, as they must be connected by a wire or radio link to a computer and attached to the speaker's body before the HCI can begin. As mentioned previously, this makes HCI impractical in many situations where it would be most helpful; e.g., when communicating with humanoid robots, or in intelligent room environments.

Although ASR is already used in several commercially available products, there are still obstacles to be overcome in making DSR commercially viable. The two major sources

of degradation in DSR are distortions, such as additive noise and reverberation, and a mismatch between *training* and *test data*, such as those introduced by speaking style or accent. In DSR scenarios, the quality of the speech provided to the recognizer has a decisive impact on system performance. This implies that speech enhancement techniques are typically required to achieve the best possible signal quality.

In the last decades, many methods have been proposed to enable ASR systems to compensate or adapt to mismatch due to interspeaker differences, articulation effects and microphone characteristics. Today, those systems work well for different users on a broad variety of applications, but only as long as the speech captured by the microphones is free of other distortions. This explains the severe performance degradation encountered in current ASR systems as soon as the microphone is moved away from the speaker's mouth. Such situations are known as *distant*, far-field or hands-free[2] speech recognition. This dramatic drop in performance occurs mainly due to three different types of distortion:

- The first is *noise*, also known as *background noise*,[3] which is any sound other than the desired speech, such as that from air conditioners, printers, machines in a factory, or speech from other speakers.
- The second distortion is *echo* and *reverberation*, which are reflections of the sound source arriving some time after the signal on the direct path.
- Other types of distortions are introduced by environmental factors such as *room modes*, the *orientation of the speaker's head*, or the *Lombard effect*.

To limit the degradation in system performance introduced by these distortions, a great deal of current research is devoted to exploiting several aspects of speech captured with far-field sensors. In DSR applications, procedures already known from conventional ASR can be adopted. For instance, *confusion network combination* is typically used with data captured with a close-talking microphone to fuse word hypotheses obtained by using various speech feature extraction schemes or even completely different ASR systems. For DSR with multiple microphone conditions, confusion network combination can be used to fuse word hypotheses from different microphones. Speech recognition with distant sensors also introduces the possibility, however, of making use of techniques that were either developed in other areas of signal processing, or that are entirely novel. It has become common in the recent past, for example, to place a *microphone array* in the speaker's vicinity, enabling the speaker's position to be determined and tracked with time. Through beamforming techniques, a microphone array can also act as a spatial filter to emphasize the speech of the desired speaker while suppressing ambient noise or simultaneous speech from other speakers. Moreover, human speech has temporal, spectral, and statistical characteristics that are very different from those possessed by other signals for which conventional beamforming techniques have been used in the past. Recent research has revealed that these characteristics can be exploited to perform more effective beamforming for speech enhancement and recognition.

[2] The latter term is misleading, inasmuch close-talking microphones are usually not held in the hand, but are mounted to the head or body of the speaker.

[3] This term is also misleading, in that the "background" could well be closer to the microphone than the "foreground" signal of interest.

1.3 System Evaluation

Quantitative measures of the quality or performance of a system are essential for making fundamental advances in the state-of-the-art. This fact is embodied in the often repeated statement, "You improve what you *measure*." In order to asses system performance, it is essential to have error metrics or objective functions at hand which are well-suited to the problem under investigation. Unfortunately, good objective functions do not exist for a broad variety of problems, on the one hand, or else cannot be directly or automatically evaluated, on the other.

Since the early 1980s, *word error rate* (WER) has emerged as the measure of first choice for determining the quality of automatically-derived speech transcriptions. As typically defined, an error in a speech transcription is of one of three types, all of which we will now describe. A *deletion* occurs when the recognizer fails to hypothesize a word that *was* spoken. An *insertion* occurs when the recognizer hypothesizes a word that was *not* spoken. A *substitution* occurs when the recognizer *misrecognizes* a word. These three errors are illustrated in the following partial hypothesis, where they are labeled with `D`, `I`, and `S`, respectively:

```
Hyp: BUT ...    WILL SELL THE CHAIN ... FOR EACH STORE SEPARATELY
Utt:     ... IT WILL SELL THE CHAIN ... OR  EACH STORE SEPARATELY
     I       D                          S
```

A more thorough discussion of word error rate is given in Section 14.1.

Even though widely accepted and used, word error rate is not without flaws. It has been argued that the equal weighting of words should be replaced by a context sensitive weighting, whereby, for example, information-bearing keywords should be assigned a higher weight than functional words or articles. Additionally, it has been asserted that word similarities should be considered. Such approaches, however, have never been widely adopted as they are more difficult to evaluate and involve subjective judgment. Moreover, these measures would raise new questions, such as how to measure the distance between words or which words are important.

Naively it could be assumed that WER would be sufficient in ASR as an objective measure. While this may be true for the user of an ASR system, it does not hold for the engineer. In fact a broad variety of additional *objective* or *cost functions* are required. These include:

- The *Mahalanobis distance*, which is used to evaluate the acoustic model.
- *Perplexity*, which is used to evaluate the language model as described in Section 7.3.1.
- *Class separability*, which is used to evaluate the feature extraction component or front-end.
- *Maximum mutual information* or *minimum phone error*, which are used during discriminate estimation of the parameters in a hidden Markov model.
- *Maximum likelihood*, which is the metric of first choice for the estimation of all system parameters.

A DSR system requires additional objective functions to cope with problems not encountered in data captured with close-talking microphones. Among these are:

- *Cross-correlation*, which is used to estimate time delays of arrival between microphone pairs as described in Section 10.1.
- *Signal-to-noise ratio*, which can be used for channel selection in a multiple-microphone data capture scenario.
- *Negentropy*, which can be used for combining the signals captured by all sensors of a microphone array.

Most of the objective functions mentioned above are useful because they show a significant correlation with WER. The performance of a system is optimized by minimizing or maximizing a suitable objective function. The way in which this optimization is conducted depends both on the objective function and the nature of the underlying model. In the best case, a closed-form solution is available, such as in the optimization of the beamforming weights as discussed in Section 13.3. In other cases, an iterative solution can be adopted, such as when optimizing the parameters of a *hidden Markov model* (HMM) as discussed in Chapter 8. In still other cases, numerical optimization algorithms must be used such as when optimization the parameters of an all-pass transform for speaker adaptation as discussed in Section 9.2.2.

To chose the appropriate objective function a number of decisions must be made (Hänsler and Schmidt 2004, sect. 4):

- What kind of information is available?
- How should the available information be used?
- How should the error be weighted by the objective function?
- Should the objective function be deterministic or stochastic?

Throughout the balance of this text, we will strive to answer these questions whenever introducing an objective function for a particular application or in a particular context. When a given objective function is better suited than another for a particular purpose, we will indicate why. As mentioned above, the reasoning typically centers around the fact that the better suited objective function is more closely correlated with word error rate.

1.4 Fields of Speech Recognition

Figure 1.7 presents several subtopics of speech recognition in general which can be associated with three different fields: automatic, robust and distant speech recognition. While some topics such as multilingual speech recognition and language modeling can be clearly assigned to one group (i.e., *automatic*) other topics such as feature extraction or adaptation cannot be uniquely assigned to a single group. A second classification of topics shown in Figure 1.7 depends on the number and type of sensors. Whereas one microphone is traditionally used for recognition, in distant recognition the traditional sensor configuration can be augmented by an entire array of microphones with known or unknown geometry. For specific tasks such as lipreading or speaker localization, additional sensor types such as video cameras can be used.

Undoubtedly, the construction of optimal DSR systems must draw on concepts from several fields, including acoustics, signal processing, pattern recognition, speaker tracking and beamforming. As has been shown in the past, all components can be optimized

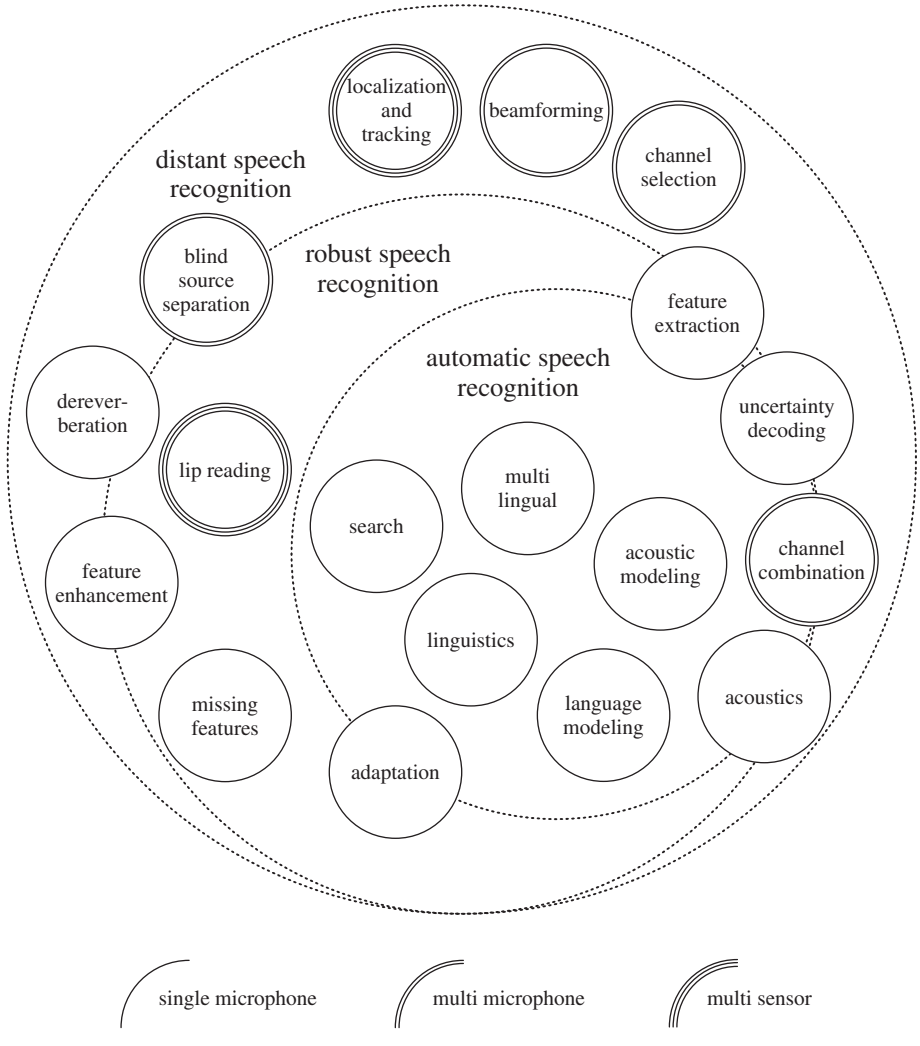

Figure 1.7 Illustration of the different fields of speech recognition: automatic, robust and distant

separately to construct a DSR system. Such an independent treatment, however, does not allow for optimal performance. Moreover, new techniques have recently emerged exploiting the complementary effects of the several components of a DSR system. These include:

- More closely coupling the feature extraction and acoustic models; e.g., by propagating the uncertainty of the feature extraction into the HMM.
- Feeding the word hypotheses produced by the DSR back to the component located earlier in the processing chain; e.g. by feature enhancement with particle filters with models for different phoneme classes.

- Replacing traditional objective functions such as signal-to-noise ratio by objective functions taking into account the acoustic model of the speech recognition system, as in maximum likelihood beamforming, or considering the particular characteristics of human speech, as in maximum negentropy beamforming.

1.5 Robust Perception

In contrast to automatic pattern recognition, human perception is very robust in the presence of distortions such as noise and reverberation. Therefore, knowledge of the mechanisms of human perception, in particular with regard to robustness, may also be useful in the development of automatic systems that must operate in difficult acoustic environments. It is interesting to note that the cognitive load for humans increases while listening in noisy environments, even when the speech remains intelligible (Kjellberg *et al*. 2007). This section presents some illustrative examples of human perceptual phenomena and robustness. We also present several technical solutions based on these phenomena which are known to improve robustness in automatic recognition.

1.5.1 A Priori Knowledge

When confronted with an ambiguous stimulus requiring a single interpretation, the human brain must rely on *a priori* knowledge and expectations. What is likely to be one of the most amazing findings about the robustness and flexibility of human perception and the use of *a priori* information is illustrated by the following sentence, which was circulated in the Internet in September 2003:

> Aoccdrnig to rscheearch at Cmabrigde uinervtisy, it deosn't mttaer waht oredr the ltteers in a wrod are, the olny ipromoetnt tihng is taht the frist and lsat ltteres are at the rghit pclae. The rset can be a tatol mses and you can sitll raed it wouthit a porbelm. Tihs is bcuseae we do not raed ervey lteter by itslef but the wrod as a wlohe.

The text is easy to read for a human inasmuch as, through reordering, the brain maps the erroneously presented characters into correct English words.

A priori knowledge is also widely used in automatic speech processing. Obvious examples are

- the statistics of speech,
- the limited number of possible phoneme combinations constrained by known words which might be further constrained by the domain,
- the word sequences follow a particular structure which can be represented as a *context free grammar* or the knowledge of successive words, represented as an *N-gram*.

1.5.2 Phonemic Restoration and Reliability

Most signals of interest, including human speech, are highly redundant. This redundancy provides for correct recognition or classification even in the event that the signal is partially

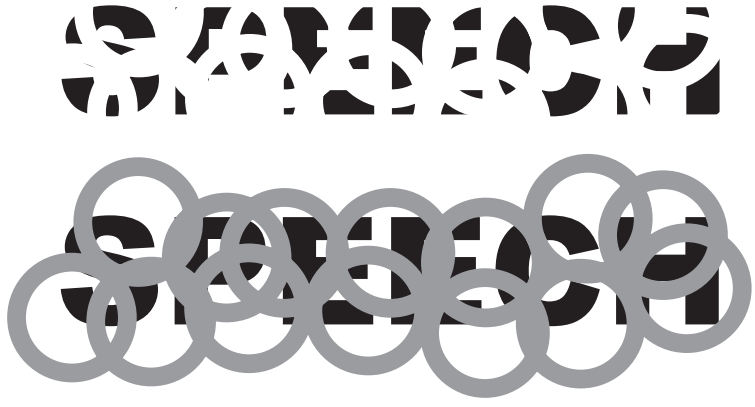

Figure 1.8 Adding a mask to the occluded portions of the top image renders the word legible, as is evident in the lower image

occluded or otherwise distorted, which implies that a significant amount of information is missing. The sophisticated capabilities of the human brain underlying robust perception were demonstrated by Fletcher (1953), who found that verbal communication between humans is possible if either the frequencies below or above 1800 Hz are filtered out. An illusory phenomenon, which clearly illustrates the robustness of the human auditory system, is known as the *phonemic restoration* effect, whereby phonetic information that is actually missing in a speech signal can be synthesized by the brain and clearly *heard* (Miller and Licklider 1950; Warren 1970). Furthermore, the knowledge of which information is distorted or missing can significantly improve perception. For example, knowledge about the occluded portion of an image can render a word readable, as is apparent upon considering Figure 1.8. Similarly, the comprehensibility of speech can be improved by adding noise (Warren *et al.* 1997).

Several problems in automatic data processing – such as occlusion – which were first investigated in the context of visual pattern recognition, are now current research topics in robust speech recognition. One can distinguish between two related approaches for coping with this problem:

- *missing feature theory*
  In missing feature theory, unreliable information is either ignored, set to some fixed nominal value, such as the global mean, or interpolated from nearby reliable information. In many cases, however, the restoration of missing features by spectral and/or temporal interpolation is less effective than simply ignoring them. The reason for this is that no processing can re-create information that has been lost as long as no additional information, such as an estimate of the noise or its propagation, is available.
- *uncertainty processing*
  In uncertainty processing, unreliable information is assumed to be unaltered, but the unreliable portion of the data is assigned less weight than the reliable portion.

1.5.3 Binaural Masking Level Difference

Even though the most obvious benefit from binaural hearing lies in source localization, other interesting effects exist: If the same signal and noise is presented to both ears with a noise level so high as to mask the signal, the signal is inaudible. Paradoxically, if either of the two ears is unable to hear the signal, it becomes once more audible. This effect is known as the *binaural masking level difference*. The binaural improvements in observing a signal in noise can be up to 20 dB (Durlach 1972). As discussed in Section 6.9.1, the binaural masking level difference can be related to spectral subtraction, wherein two input signals, one containing both the desired signal along with noise, and the second containing only the noise, are present. A closely related effect is the so-called *cocktail party effect* (Handel 1989), which describes the capacity of humans to suppress undesired sounds, such as the babble during a cocktail party, and concentrate on the desired signal, such as the voice of a conversation partner.

1.5.4 Multi-Microphone Processing

The use of multiple microphones is motivated by nature, in which two ears have been shown to enhance speech understanding as well as acoustic source localization. This effect is even further extended for a group of people, where one person could not understand some words, a person next to the first might have and together they are able to understand more than independent of each other.

Similarly, different tiers in a speech recognition system, which are derived either from different channels (e.g., microphones at different locations or visual observations) or from the variance in the recognition system itself, produce different recognition results. An appropriate combination of the different tiers can improve recognition performance. The degree of success depends on

- the variance of the information provided by the different tiers,
- the quality and reliability of the different tiers and
- the method used to combine the different tiers.

In automatic speech recognition, the different tiers can be combined at various stages of the recognition system providing different advantages and disadvantages:

- *signal combination*
  Signal-based algorithms, such as *beamforming*, exploit the spatial diversity resulting from the fact that the desired and interfering signal sources are in practice located at different points in space. These approaches assume that the time delays of the signals between different microphone pairs are known or can be reliably estimated. The spatial diversity can then be exploited by suppressing signals coming from directions other than that of the desired source.
- *feature combination*
  These algorithms concatenate features derived by different feature extraction methods to form a new feature vector. In such an approach, it is a common practice to reduce the number of features by principal component analysis or linear discriminant analysis.

While such algorithms are simple to implement, they suffer in performance if the different streams are not perfectly synchronized.

- *word and lattice combination*
  Those algorithms, such as *recognizer output voting error reduction* (ROVER) and confusion network combination, combine the information of the recognition output which can be represented as a first best, N-best or lattice word sequence and might be augmented with a confidence score for each word.

In the following we present some examples where different tiers have been successfully combined: Stolcke *et al.* (2005) used two different front-ends, mel-frequency cepstral coefficients and features derived from perceptual linear prediction, for cross-adaptation and system combination via confusion networks. Both of these features are described in Chapter 5. Yu *et al.* (2004) demonstrated, on a Chinese ASR system, that two different kinds of models, one on phonemes, the other on semi-syllables, can be combined to good effect. Lamel and Gauvain (2005) combined systems trained with different phoneme sets using ROVER. Siohan *et al.* (2005) combined randomized decision trees. Stüker *et al.* (2006) showed that a combination of four systems – two different phoneme sets with two feature extraction strategies – leads to additional improvements over the combination of two different phoneme sets or two front-ends. Stüker *et al.* also found that combining two systems, where both the phoneme set and front-ends are altered, leads to improved recognition accuracy compared to changing only the phoneme set or only the front-end. This fact follows from the increased variance between the two different channels to be combined. The previous systems have combined different tiers using only a single channel combination technique. Wölfel *et al.* (2006) demonstrated that a hybrid approach combining the different tiers, derived from different microphones, at different stages in a distant speech recognition system leads to additional improvements over a single combination approach. In particular Wölfel *et al.* achieved fewer recognition errors by using a combination of beamforming and confusion network.

1.5.5 Multiple Sources by Different Modalities

Given that it often happens that no single modality is powerful enough to provide correct classification, one of the key issues in robust human perception is the efficient merging of different input modalities, such as audio and vision, to render a stimulus intelligible (Ernst and Bülthoff 2004; Jacobs 2002). An illustrative example demonstrating the multimodality of speech perception is the *McGurk effect*[4] (McGurk and MacDonald 1976), which is experienced when contrary audiovisual information is presented to human subjects. To wit, a video presenting a visual /ga/ combined with an audio /ba/ will be perceived by 98% of adults as the syllable /da/. This effect exists not only for single syllables, but can alter the perception of entire spoken utterances, as was confirmed by a study about witness testimony (Wright and Wareham 2005). It is interesting to note that awareness of the effect does not change the perception. This stands in stark contrast to certain optical illusions, which are destroyed as soon as the subject is aware of the deception.

[4] This is often referred to as the McGurk–MacDonald effect.

Humans follow two different strategies to combine information:

- *maximizing information (sensor combination)*
  If the different modalities are complementary, the various pieces of information about an object are combined to maximize the knowledge about the particular observation.
  For example, consider a three-dimensional object, the correct recognition of which is dependent upon the orientation of the object to the observer. Without rotating the object, vision provides only two-dimensional information about the object, while the haptic[5] input provides the missing three-dimensional information (Newell 2001).
- *reducing variance (sensor integration)*
  If different modalities overlap, the variance of the information is reduced. Under the independence and Gaussian assumption of the noise, the estimate with the lowest variance is identical to the maximum likelihood estimate.
  One example of the integration of audio and video information for localization supporting the reduction in variance theory is given by Alais and Burr (2004).

Two prominent technical implementations of sensor fusion are audio-visual speaker tracking, which will be presented in Section 10.4, and audio-visual speech recognition. A good overview paper of the latter is by Potamianos *et al*. (2004).

1.6 Organizations, Conferences and Journals

Like all other well-established scientific disciplines, the fields of speech processing and recognition have founded and fostered an elaborate network of conferences and publications. Such networks are critical for promoting and disseminating scientific progress in the field. The most important organizations that plan and hold such conferences on speech processing and publish scholarly journals are listed in Table 1.1.

At conferences and in their associated proceedings the most recent advances in the state-of-the-art are reported, discussed, and frequently lead to further advances. Several major conferences take place every year or every other year. These conferences are listed in Table 1.2. The principal advantage of conferences is that they provide a venue for

Table 1.1 Organizations promoting research in speech processing and recognition

Abbreviation	Full Name
IEEE	Institute of Electrical and Electronics Engineers
ISCA	International Speech Communication Association former European Speech Communication Association (ESCA)
EURASIP	European Association for Signal Processing
ASA	Acoustical Society of America
ASJ	Acoustical Society of Japan
EAA	European Acoustics Association

[5] Haptic phenomena pertain to the sense of touch.

Table 1.2 Speech processing and recognition conferences

Abbreviation	Full Name
ICASSP	International Conference on Acoustics, Speech, and Signal Processing by IEEE
Interspeech	ISCA conference; previous Eurospeech and International Conference on Spoken Language Processing (ICSLP)
ASRU	Automatic Speech Recognition and Understanding by IEEE
EUSIPCO	European Signal Processing Conference by EURASIP
HSCMA	Hands-free Speech Communication and Microphone Arrays
WASPAA	Workshop on Applications of Signal Processing to Audio and Acoustics
IWAENC	International Workshop on Acoustic Echo and Noise Control
ISCSLP	International Symposium on Chinese Spoken Language Processing
ICMI	International Conference on Multimodal Interfaces
MLMI	Machine Learning for Multimodal Interaction
HLT	Human Language Technology

the most recent advances to be reported. The disadvantage of conferences is that the process of peer review by which the papers to be presented and published are chosen is on an extremely tight time schedule. Each submission is either accepted or rejected, with no time allowed for discussion with or clarification from the authors. In addition to the scientific papers themselves, conferences offer a venue for presentations, expert panel discussions, keynote speeches and exhibits, all of which foster further scientific progress in speech processing and recognition. Information about individual conferences is typically disseminated in the Internet. For example, to learn about the *Workshop on Applications of Signal Processing to Audio and Acoustics*, which is to be held in 2009, it is only necessary to type `waspaa 2009` into an Internet search window.

Journals differ from conferences in two ways. Firstly, a journal offers no chance for the scientific community to gather regularly at a specific place and time to present and discuss recent research. Secondly and more importantly, the process of peer review for an article submitted for publication in a journal is far more stringent than that for any conference. Because there is no fixed time schedule for publication, the reviewers for a journal can place far more demands on authors prior to publication. They can, for example, request more graphs or figures, more experiments, further citations to other scientific work, not to mention improvements in English usage and overall quality of presentation. While all of this means that greater time and effort must be devoted to the preparation and revision of a journal publication, it is also the primary advantage of journals with respect to conferences. The dialogue that ensues between the authors and reviewers of a journal publication is the very core of the scientific process. Through the succession of assertion, rebuttal, and counter assertion, non-novel claims are identified and withdrawn, unjustifiable claims are either eliminated or modified, while the arguments for justifiable claims are strengthened and clarified. Moreover, through the act of publishing a journal article and the associated dialogue, both authors and reviewers typically learn much they had not previously known. Table 1.3 lists several journals which cover topics presented in this book and which are recognized by academia and industry alike.

Table 1.3 Speech processing and recognition journals

Abbreviation	Full name
SP	*IEEE Transactions on Signal Processing*
ASLP	*IEEE Transactions on Audio, Speech and Language Processing* former *IEEE Transactions on Speech and Audio Processing* (SAP)
ASSP	*IEEE Transactions on Acoustics, Speech and Signal Processing*
SPL	*IEEE Signal Processing Letters*
SPM	*IEEE Signal Processing Magazine*
CSL	*Computer Speech and Language* by Elsevier
ASA	*Journal of the Acoustic Society of America*
SP	*EURASIP Journal on Signal Processing*
AdvSP	*EURASIP Journal on Advances in Signal Processing*
SC	*EURASIP and ISCA Journal on Speech Communication* published by Elsevier
AppSP	*EURASIP Journal on Applied Signal Processing*
ASMP	*EURASIP Journal on Audio, Speech and Music Processing*

An updated list of conferences, including a calendar of upcoming events, and journals can be found on the companion website of this book at

`http://www.distant-speech-recognition.org`

1.7 Useful Tools, Data Resources and Evaluation Campaigns

A broad number of commercial and non-commercial tools are available for the processing, analysis and recognition of speech. An extensive and updated list of such tools can be found on the companion website of this book.

The right data or corpora is essential for training and testing various speech processing, enhancement and recognition algorithms. This follows from the fact that the quality of the acoustic and language models are determined in large part by the amount of available training data, and the similarity between the data used for training and testing. As collecting and transcribing appropriate data is time-consuming and expensive, and as reporting WER reductions on "private" data makes the direct comparison of techniques and systems difficult or impossible, it is highly worth-while to report experimental results on publicly available speech corpora whenever possible. The goal of evaluation campaigns, such as the *Rich Transcription* (RT) evaluation staged periodically by the US *National Institute of Standards and Technologies* (NIST), is to evaluate and to compare different speech recognition systems and the techniques on which they are based. Such evaluations are essential in order to assess not only the progress of individual systems, but also that of the field as a whole. Possible data sources and evaluation campaigns are listed on the website mentioned previously.

1.8 Organization of this Book

Our aim in writing this book was to provide in a single volume an exposition of the theory behind each component of a complete DSR system. We now summarize the remaining

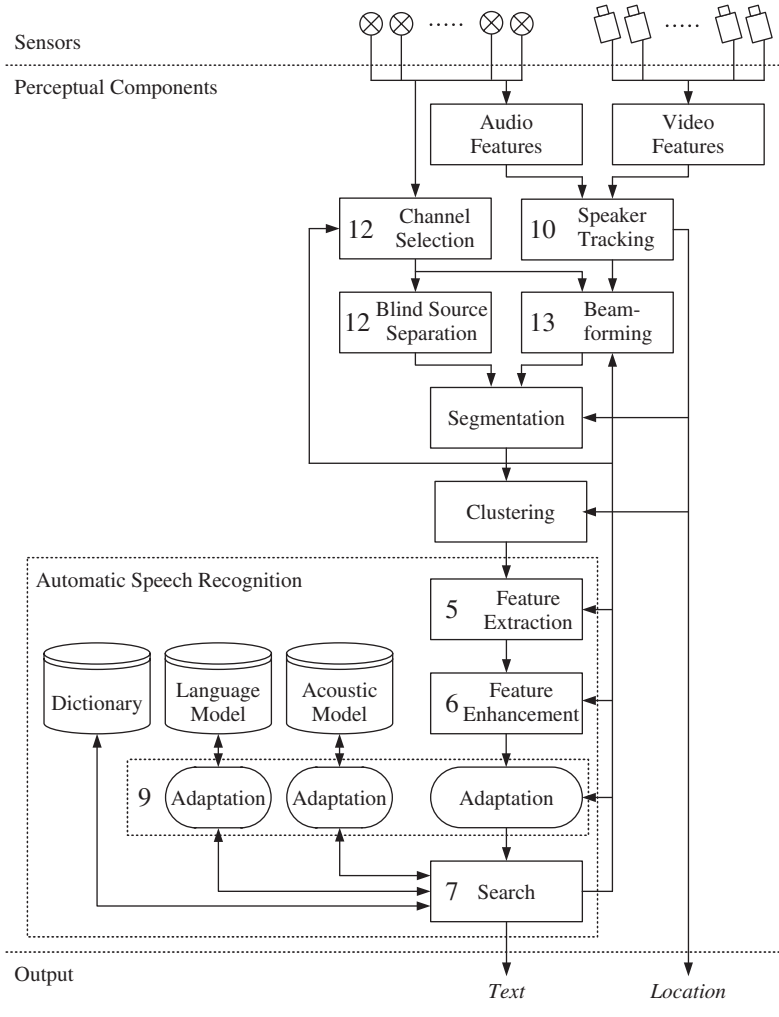

Figure 1.9 Architecture of a distant speech recognition system. The gray numbers indicate the corresponding chapter of this book

contents of this volume in order to briefly illustrate both the narrative thread that underlies this work, as well as the interrelations among the several chapters. In particular, we will emphasize how the development of each chapter is prefigured by and builds upon that of the preceding chapters. Figure 1.9 provides a high-level overview of a DSR system following the signal flow through the several components. The gray number on each individual component indicates the corresponding chapter in this book. The chapters not shown in the figure, in particular Chapters 2, 3, 4, 8 and 11, present material necessary to support the development in the other chapters: The fundamentals of sound propagation and acoustics are presented in Chapter 2, as are the basics of speech production. Chapter 3 presents linear filtering techniques that are used throughout the text. Chapter 4 presents the theory of Bayesian filters, which will later be applied both for speech feature enhancement

in Chapter 6 and speaker tracking in Chapter 10. Chapter 8 discusses how the parameters of a HMM can be reliably estimated based on the use of transcribed acoustic data. Such a HMM is an essential component of most current DSR systems, in that it extracts word hypotheses from the final waveform produced by the other components of the system. Chapter 11 provides a discussion of digital filter banks, which, as discussed in Chapter 13, are an important component of a beamformer. Finally, Chapter 14 reports experimental results indicating the effectiveness of the algorithms described throughout this volume.

Speech, like any sound, is the propagation of pressure waves through air or any other liquid. A DSR system extracts from such pressure waves hypotheses of the phonetic units and words uttered by a speaker. Hence, it is worth-while to understand the physics of sound propagation, as well as how the spectral and temporal characteristics of speech are altered when it is captured by far-field sensors in realistic acoustic environments. These topics are considered in Chapter 2. This chapter also presents the characteristics and properties of the human auditory system. Knowledge of the latter is useful, inasmuch as experience has shown that many insights gained from studying the human auditory system have been successfully applied to improve the performance of *automatic* speech recognition systems.

In signal processing, the term *filter* refers to an algorithm which extracts a desired signal from an input signal corrupted by noise or other distortions. A filter can also be used to modify the spectral or temporal characteristics of a signal in some advantageous way. Therefore, filtering techniques are powerful tools for speech signal processing and distant recognition. Chapter 3 provides a review of the basics of digital signal processing, including a short introduction to linear time-invariant systems, the Fourier and z-transforms, as well as the effects of sampling and reconstruction. Next there is a presentation of the discrete Fourier transform and its use for the implementation of linear time-invariant systems, which is followed by a description of the short-time Fourier transform. The contents of this chapter will be referred to extensively in Chapter 5 on speech feature extraction, as well as in Chapter 11 on digital filter banks.

Many problems in science and engineering can be formulated as the estimation of some *state*, which cannot be observed directly, based on a series of features or observations, which can be directly observed. The observations are often corrupted by distortions such as noise or reverberation. Such problems can be solved with one of a number of Bayesian filters, all of which estimate an unobservable state given a series of observations. Chapter 4 first formulates the general problem to be solved by a Bayesian filter, namely, tracking the likelihood of the state as it evolves in time as conditioned on a sequence of observations. Thereafter, it presents several different solutions to this general problem, including the classic Kalman filter and its variants, as well as the class of particle filters, which have much more recently appeared in the literature. The theory of Bayesian filters will be applied in Chapter 6 to the task of enhancing speech features that have been corrupted by noise, reverberation or both. A second application, that of tracking the physical position of a speaker based on the signals captured with the elements of a microphone array, will be discussed in Chapter 10.

Automatic recognition requires that the speech waveform is processed so as to produce feature vectors of a relatively small dimension. This reduction in dimensionality is necessary in order to avoid wasting parameters modeling characteristics of the signal which are irrelevant for classification. The transformation of the input data into a set of dimension-reduced features is called speech feature extraction, acoustic preprocessing

or front-end processing. As explained in Chapter 5, feature extraction in the context of DSR systems aims to preserve the information needed to distinguish between phonetic classes, while being invariant to other factors. The latter include speaker differences, such as accent, emotion or speaking rate, as well as environmental distortions such as background noise, channel differences, or reverberation.

The principle underlying speech feature enhancement, the topic of Chapter 6, is the estimation of the original features of the clean speech from a corrupted signal. Usually the enhancement takes place either in the power, logarithmic spectral or cepstral domain. The prerequisite for such techniques is that the noise or the impulse response is known or can be reliably estimated in the cases of noise or channel distortion, respectively. In many applications only a single channel is available and therefore the noise estimate must be inferred directly from the noise-corrupted signal. A simple method for accomplishing this separates the signal into speech and non-speech regions, so that the noise spectrum can be estimated from those regions containing no speech. Such simple techniques, however, are not able to cope well with non-stationary distortions. Hence, more advanced algorithms capable of actively tracking changes in the noise and channel distortions are the main focus of Chapter 6.

As discussed in Chapter 7, search is the process by which a statistical ASR system finds the most likely word sequence conditioned on a sequence of acoustic observations. The search process can be posed as that of finding the shortest path through a search graph. The construction of such a search graph requires several knowledge sources, namely, a language model, a word lexicon, and a HMM, as well as an acoustic model to evaluate the likelihoods of the acoustic features extracted from the speech to be recognized. Moreover, inasmuch as all human speech is affected by coarticulation, a decision tree for representing context dependency is required in order to achieve state-of-the-art performance. The representation of these knowledge sources as weighted finite-state transducers is also presented in Chapter 7, as are weighted composition and a set of equivalence transformations, including determinization, minimization, and epsilon removal. These algorithms enable the knowledge sources to be combined into a single search graph, which can then be optimized to provide maximal search efficiency.

All ASR systems based on the HMM contain an enormous number of free parameters. In order to train these free parameters, dozens if not hundreds or even thousands of hours of transcribed acoustic data are required. Parameter estimation can then be performed according to either a maximum likelihood criterion or one of several discriminative criteria such as maximum mutual information or minimum phone error. Algorithms for efficiently estimating the parameters of a HMM are the subjects of Chapter 8. Included among these are a discussion of the well-known expectation-maximization algorithm, with which maximum likelihood estimation of HMM parameters is almost invariably performed. Several discriminative optimization criteria, namely, maximum mutual information, and minimum word and phone error are also described.

The unique characteristics of the voice of a particular speaker are what allow a person calling on the telephone to be identified as soon as a few syllables have been spoken. These characteristics include fundamental frequency, speaking rate, and accent, among others. While lending each voice its own individuality and charm, such characteristics are a hindrance to automatic recognition, inasmuch as they introduce variability in the speech that is of no use in distinguishing between different words. To enhance the performance

of an ASR system that must function well for any speaker as well as different acoustic environments, various transformations are typically applied either to the features, the means and covariances of the acoustic model, or to both. The body of techniques used to estimate and apply such transformations fall under the rubrik *feature and model adaptation* and comprise the subject matter of Chapter 9.

While a recognition engine is needed to convert waveforms into word hypotheses, the speech recognizer by itself is not the only component of a distant recognition system. In Chapter 10, we introduce an important supporting technology required for a complete DSR system, namely, algorithms for determining the physical positions of one or more speakers in a room, and tracking changes in these positions with time. Speaker localization and tracking – whether based on acoustic features, video features, or both – are important technologies, because the beamforming algorithms discussed in Chapter 13 all assume that the position of the desired speaker is *known*. Moreover, the accuracy of a speaker tracking system has a very significant influence on the recognition accuracy of the entire system.

Chapter 11 discusses digital filter banks, which are arrays of bandpass filters that separate an input signal into many narrowband components. As mentioned previously, frequent reference will be made to such filter banks in Chapter 13 during the discussion of beamforming. The optimal design of such filter banks has a critical effect on the final system accuracy.

Blind source separation (BSS) and *independent component analysis* (ICA) are terms used to describe classes of techniques by which signals from multiple sensors may be combined into one signal. As presented in Chapter 12, this class of methods is known as *blind* because neither the relative positions of the sensors, nor the position of the sources are assumed to be known. Rather, BSS algorithms attempt to separate different sources based only on their temporal, spectral, or statistical characteristics. Most information-bearing signals are non-Gaussian, and this fact is extremely useful in separating signals based only on their statistical characteristics. Hence, the primary assumption of ICA is that interesting signals are *not* Gaussian signals. Several optimization criteria that are typically applied in the ICA field include kurtosis, negentropy, and mutual information. While mutual information can be calculated for both Gaussian and non-Gaussian random variables alike, kurtosis and negentropy are only meaningful for non-Gaussian signals. Many algorithms for blind source separation, dispense with the assumption of non-Gaussianity and instead attempt to separate signals on the basis of their non-stationarity or non-whiteness. Insights from the fields of BSS and ICA will also be applied to good effect in Chapter 13 for developing novel beamforming algorithms.

Chapter 13 presents a class of techniques, known collectively as beamforming, by which signals from several sensors can be combined to emphasize a desired source and to suppress all other noise and interference. Beamforming begins with the assumption that the positions of all sensors are known, and that the positions of the desired sources are known or can be estimated. The simplest of beamforming algorithms, the delay-and-sum beamformer, uses only this geometrical knowledge to combine the signals from several sensors. More sophisticated adaptive beamformers attempt to minimize the total output power of an array of sensors under a constraint that the desired source must be unattenuated. Recent research has revealed that such optimization criteria used in conventional array processing are not optimal for acoustic beamforming applications. Hence, Chapter

13 also presents several nonconventional beamforming algorithms based on optimization criteria – such as mutual information, kurtosis, and negentropy – that are typically used in the fields of BSS or ICA.

In the final chapter of this volume we present the results of performance evaluations of the algorithms described here on several DSR tasks. These include an evaluation of the speaker tracking component in isolation from the rest of the DSR system. In Chapter 14, we present results illustrating the effectiveness of single-channel speech feature enhancement based on particle filters. Also included are experimental results for systems based on beamforming for both single distant speakers, as well as two simultaneously active speakers. In addition, we present results illustrating the importance of selecting a filter bank suitable for adaptive filtering and beamforming when designing a complete DSR system.

A note about the brevity of the chapters mentioned above is perhaps now in order. To wit, each of these chapters might easily be expanded into a book much larger than the present volume. Indeed, such books are readily available on sound propagation, digital signal processing, Bayesian filtering, speech feature extraction, HMM parameter estimation, finite-state automata, blind source separation, and beamforming using conventional criteria. Our goal in writing this work, however, was to create an accessible description of all the components of a DSR system required to transform sound waves into word hypotheses, including metrics for gauging the efficacy of such a system. Hence, judicious selection of the topics covered along with concise presentation were the criteria that guided the choice of every word written here. We have, however, been at pains to provide references to lengthier specialized works where applicable – as well as references to the most relevant contributions in the literature – for those desiring a deeper knowledge of the field. Indeed, this volume is intended as a starting point for such wider exploration.

1.9 Principal Symbols used Throughout the Book

This section defines principal symbols which are used throughout the book. Due to the numerous variables each chapter presents an individual list of principal symbols which is specific for the particular chapter.

Symbol	Description
$a, b, c, \ldots$	variables
$A, B, C, \ldots$	constants
a, b, c, A, B, C, ...	units
$\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$	vectors
$\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$	matrices
$\mathbf{I}$	unity matrix
j	imaginary number, $\sqrt{-1}$
$.^*$	complex conjungate

Symbol	Description
$.^T$	transpose operator
$.^H$	Hermetian operator
$\cdot_{1:K}$	sequence from 1 to K
∇^2	Laplace operator
$\bar{\cdot}$	average
$\tilde{\cdot}$	warped frequency
$\hat{\cdot}$	estimate
$\%$	modulo
λ	Lagrange multiplier
$(\cdot)^+$	pseudoinverse of $(\cdot)$
$\mathcal{E}\{\cdot\}$	expectation value
$/\cdot/$	denote a phoneme
$[\cdot]$	denote a phone
$\lvert\cdot\rvert$	absolute (scalar) or determinant (matrix)
$\boldsymbol{\mu}$	mean
$\boldsymbol{\Sigma}$	covariance matrix
$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\forall$	for all
$*$	convolution
δ	Dirac impulse
$\mathcal{O}$	big O notation also called Landau notation
$\mathbb{C}$	complex number
$\mathbb{N}$	set of natural numbers
$\mathbb{N}_0$	set of non-negative natural numbers including zero
$\mathbb{R}$	real number
$\mathbb{R}^+$	non-negative real number
$\mathbb{Z}$	integer number
$\mathbb{Z}^+$	non-negative integer number
$\operatorname{sinc}(z)$	$\triangleq \begin{cases} 1, & \text{for } z = 0, \\ \sin(z)/z, & \text{otherwise} \end{cases}$

1.10 Units used Throughout the Book

This section defines units that are consistently defined throughout the book.

Units	Description
Hz	Herz
J	Joule
K	Kelvin
Pa	Pascal
SPL	sound pressure level
Vs/m^2	Tesla
W	Watt
$^{\circ}C$	degree Celsius
dB	decibel
kg	kilogram
m	meter
m^2	square meter
m^3	cubic meter
m/s	velocity
s	second